ISDS 2012 Conference Abstracts

Tweeting Fever: Are Tweet Extracts a Valid Surrogate Data Source for Dengue Fever?

Jacqueline S. Coberly*[1], Clayton R. Fink[1], Eugene Elbert[1], In-Kyu Yoon[2], John M. Velasco[2], Agnes Tomayo[2], V. Roque[3], S. Ygano[4], Durinda Macasoco[4] and Sheri Lewis[3]

[1]The Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA; [2]Armed Forces Research Institute for Medical Research, Bangkok, Thailand; [3]National Epidemiology Center, Manila, Philippines; [4]Cebu City Health Office, Cebu City, Philippines

Objective

To determine whether Twitter data contains information on dengue-like illness and whether the temporal trend of such data correlates with the incidence dengue or dengue-like illness as identified by city and national health authorities.

Introduction

Dengue fever is a major cause of morbidity and mortality in the Republic of the Philippines (RP) and across the world. Early identification of geographic outbreaks can help target intervention campaigns and mitigate the severity of outbreaks. Electronic disease surveillance can improve early identification but, in most dengue endemic areas data pre-existing digital data are not available for such systems. Data must be collected and digitized specifically for electronic disease surveillance. Twitter, however, is heavily used in these areas; for example, the RP is among the top 20 producers of tweets in the world. If social media could be used as a surrogate data source for electronic disease surveillance, it would provide an inexpensive pre-digitized data source for resource-limited countries. This study investigates whether Twitter extracts can be used effectively as a surrogate data source to monitor changes in the temporal trend of dengue fever in Cebu City and the National Capitol Region surrounding Manila (NCR) in the RP.

Methods

We obtained two sources of ground truth incidence for dengue. The first was daily dengue fever incidence for Cebu City and the NCR taken from the Philippines Integrated Disease Surveillance and Response System (PIDSR). The second ground truth source was fever incidence from Cebu City for 2011. The Cebu City Health Office (CCHO) has monitored fever incidence as a surrogate for dengue fever since the 1980s. Tweets from Cebu City, and the NCR were collected prospectively thru Twitter's public application program interface. The Cebu City fever ground truth data set was smoothed with a seven day moving average to facilitate comparison to the PIDSR and Twitter data. A vocabulary of words and phrases describing fever and dengue fever in the tweets collected were identified and used to mark relevant tweets. A subset of these 'fever' tweets that mentioned fever related to a medical situation were identified. The incidence and the temporal pattern of these medically-relevant tweets were compared with the incidence and pattern of fever and dengue fever in the two ground truth data sets. Pearson correlation coefficient was used to compare the correlation among the different data sets. Noted lag periods were adjusted by moving the data in time and re-computing the correlation coefficient.

Results

26,023,103 tweets were collected from the two geographic regions: 10,303,366 from Cebu City and 15,719,767 tweets from the NCR. 8,814 (0.02%) Tweets contained the word fever and 4099 (0.01% of total) mentioned fever in a medically-relevant context, for example. "…I have a fever…" vs. "…football fever…." The medically-relevant tweets were compared with both ground truth data sets. The correlation between the Tweets and each of the incidence data sets is shown below.

Conclusions

Tweets containing medically-relevant fever references were correlated ($p<0.0001$) with both fever and dengue fever incidence in the ground truth data sets. The signal indicating fever in the medically-related tweets led the incidence data significantly: by 6 days for the Cebu City fever incidence; and by 12 days for the PIDSR dengue fever incidence. Temporal adjustment to account for observed lag periods increased the correlation coefficient by about one-third in both cases. This was a limited pilot study, but it suggests that Twitter extracts may provide a valid and timely surrogate data source to monitor dengue fever in this population. Further study of the correlation of Twitter and dengue in other areas, and of Twitter with other illnesses is warranted.

Table 1: Correlation between Twitter Extracts and Fever & Dengue Fever Incidence Data Sets

	Pearson Correlation Coefficients	
Incidence Data	Raw Data	Temporally Adjusted
Twitter vs. PIDSR Dengue	0.629*	0.829*‡
Twitter vs. CCHO Fever	0.575*	0.769*†

* $p<0.0001$
† Twitter shifted right by 6 days
‡ Twitter shifted right by 12 days

Keywords

Dengue; Social Media; Twitter

*Jacqueline S. Coberly
E-mail: jacqueline.coberly@jhuapl.edu